DOCUMENT RESUME

ED 372 086 TM 021 748

AUTHOR Gipps, Caroline V.
TITLE Quality Assurance in Teachers' Assessment.
PUB DATE Apr 94
NOTE 24p.; Paper presented at the Annual Meetings of both the American Educational Research Association (New Orleans, LA, April 4-8, 1994) and the British Educational Research Association.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; *Educational Assessment; Elementary Secondary Education; *Evaluation Methods; Feedback; Foreign Countries; Formative Evaluation; Generalizability Theory; Models; Statistical Analysis; *Student Evaluation; Teacher Expectations of Students; *Teacher Made Tests; *Test Construction; Test Reliability; Test Use; Test Validity
IDENTIFIERS *Moderation; Panel Consensus Technique; *Quality Assurance; United Kingdom

ABSTRACT

The teacher assessment that is the subject of this paper is an essentially informal activity. The teacher assesses the student by posing questions, observing activities, and evaluating work in a planned or ad hoc way. The information obtained may be partial or fragmented, but repeating such assessments over time will allow the buildup of a solid and broadly based understanding of student attainment. Such assessment can certainly be formative. Examination of the assessment procedures of teachers preparing the teacher-assessment element of the British National Assessment suggests that teachers may be categorized by their approach to assessment as intuitives, evidence gatherers, and systematic planners. Issues of reliability and validity are discussed, and the moderation of teachers' assessments by the common or consensus judgments of a group or panel of teachers or experts is reviewed. Such moderation is a feature of the British National Assessment. While assessment that is internal to the school is more professionally rewarding to the teacher in terms of enhancing teaching and learning and is more valid, assessment that is to be used outside the school must be more demonstrably comparable across teachers, tasks, and pupils. Statistical moderation and other forms of quality assurance can help ensure this. (Contains 23 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 372 086

AERA Conference 1994
New Orleans

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
CAROLINE GIPPS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

Quality Assurance in Teachers' Assessment

Caroline V Gipps
University of London Institute of Education
20 Bedford Way
LONDON WC1H 0AL
UK

Paper presented in the symposium
Enhancing Quality in Student Assessment
British Educational Research Association (BERA)

To appear in
Assuring Quality in Assessment
Edited by Wynne Harlen
to be published by Paul Chapman, June 1994

BEST COPY AVAILABLE

TM021748

Quality Assurance in Teachers' Assessment

Introduction

Assessments which teachers make of pupils' attainment and performance are called variously teacher assessment (although in the US this refers to assessment of teachers), school based assessment and formative assessment. Formative assessment is best viewed as a subset of teacher assessment while teacher assessment itself can be summative as well as formative. In the UK there is a history of using teachers' assessment of pupils' work in Records of Achievement, public exams at 16, and now in the National Curriculum Assessment programme. Teacher assessment may be of a required project or task, as with the assessed course-work element of our GCSE exam. In this situation the task, marking scheme and criteria are specified and the results are subjected to some form of moderation. But teacher assessment may also be, as it were, free-floating, without the constraints of a specified task but, as with National Curriculum Assessment, with assessment criteria.

Teacher assessment with which I am concerned in this paper is essentially an informal activity: the teacher may pose questions, observe activities, evaluate pupils' work in a planned and systematic or ad hoc way. The information which the teacher thus obtains may then be partial or fragmentary; it will not at the time allow the teacher to make a firm evaluation of the pupils' competence in reading, for example, or understanding of a mathematical process. But repeated assessment of this sort, over a period of time, and in a range of contexts will allow the teacher to build up a solid and broadly-based understanding of the pupil's attainment. Because of these characteristics teacher assessment may be seen as having high validity (see Harlen's paper). If the teacher assessment is used for formative purposes which then results in improved learning then the assessment can be said to have consequential validity, ie it has the consequences expected/required of it. If the assessment has sampled broadly across the domain and

in depth within it then the assessment is likely to be generalisable (within that domain), since the teacher's evaluation of the pupil's ability to read at a certain level or to be able to manipulate single digits, will be based on a broad sample of tasks and assessments. An external test, on the other hand, will provide more limited information based as it is on a one-off occasion covering a limited sample of tasks.

Formative Assessment

Formative assessment involves using assessment information to feed back into the teaching/learning process; some believe that assessment is only truly formative if it involves the pupil, others that it can be a process which involves only the teacher who feeds back into curriculum planning. The rationale of formative assessment is linked with the constructivist model of learning. In this model it is important to understand what the child knows and how she articulates it in order to develop her knowledge and understanding. In this model it is learning with understanding which counts and to this end information about existing ideas and skills is essential. Work in psychology and learning tells us similarly that for effective learning the task must be matched to the child's current level of understanding (Gipps, 1992a) and either pitched at that level to provide practice or slightly higher in order to extend and develop the child's skills. If the new task is much too easy the child can become bored, if much too difficult the child can become de-motivated. Assessment to find out what and how children know is thus part of good teaching practice and in helping the teacher to decide what and how to teach next is formative assessment. However, if it is to be really fruitful it seems that the pupil must also be involved , since teachers need to explain to pupils what they need to do to improve their work or the next steps in the learning process.

Sadler (1989) conceptualises formative assessment as being concerned with how judgements about the quality of students' responses can be used to shape and improve their competence by short-circuiting the randomness and inefficiency of trial-and-error

learning. The key difference between formative assessment and summative assessment is not timing, but purpose and effect: assessments made during the course of a unit or session may be used for summative or grading purposes rather than for truly formative purposes.

In Sadler's classic paper (1989) formative assessment is connected with feedback and for him feedback to teacher and pupil are separated:

> 'Teachers use feedback to make programmatic decisions with respect to readiness, diagnosis and remediation. Students use it to monitor the strengths and weaknesses of their performances, so that aspects associated with success or high quality can be recognised and reinforced, and unsatisfactory aspects modified or improved'.
>
> (Sadler 1989 p120)

Sadler's work in theorising formative assessment stems from the 'common but puzzling' observation that even when teachers give students valid and reliable judgements about their work improvement does not necessarily follow. In order for the student to improve s/he must have: a notion of the desired standard or goal, be able to compare the actual performance with the desired performance and to engage in appropriate action to close the gap between the two. Feedback from the teacher, which helps the student with the second of these stages, needs to be of the kind and detail which tells the student what to do to improve; the use of grades or 'good, 7/10' marking cannot do this. Grades may in fact shift attention away from the criteria and be counter-productive for formative purposes. In Sadler's model, grades do not count as feedback: information fed back to the student is only feedback when it can be used to close the gap.

A key aspect of formative assessment, and an indispensable condition for improvement, is that the student comes to hold a notion of the standard or desired quality similar to that of the teacher, is able to monitor the quality of what is being produced at the time of production, and is able to regulate their work appropriately.

When the student reaches this stage the process is referred to as self-monitoring (rather than feedback from the teacher). Competent learners are those who self-monitor their work, although this does not mean that the need for feedback from the teacher decreases: such feedback will continue to be necessary whenever a new subject, standard or criterion is introduced.

Teacher Assessment in Practice

In our study of teachers' assessment practice at primary school level as part of the introduction of the national assessment programme, we spent a considerable amount of time trying to understand how teachers made their assessments for the Teacher Assessment (TA) element of National Assessment (reported at AERA last year and now published as McCallum et al 1993*). While infant teachers had, prior to the introduction of national assessment, made informal assessments in the basic skills in order to write holistic descriptions of pupil progress for parents, assessing in relation to tightly specified criteria at different levels was a completely new requirement. Most of the resources for the development of National Assessment went into producing test materials with little support for teacher assessment or training. Given that teachers had little preparation of support by way of a model for TA, it was not surprising that we found they adopted a range of procedures.

* Ref ESRC grant no: R00023 2192

Models of teacher assessment

We grouped teachers' approaches into three models called Intuitives, Evidence Gatherers and Systematic Planners.

These models, though describing teachers' practice in assessment, link in with teachers' (implicit) views of learning; they also link with different attitudes and approaches to criterion-referenced assessment and formative assessment.

The Intuitives rely on their memory in making and recording assessment so that there is a lack of observable TA. They do not refer to Statements of Attainment, they do not take notes, they reject systematic recorded assessment as too formal and structured an approach. Their assessment style is essentially intuitive: only the teacher can assess the child, assessment is built on close, all-round knowledge of children. This group of teachers broke down into two sub-groups. The first, the Children's Needs Ideologists, have an exploratory or 'scaffolded' view of learning, in which they provide a stimulating environment and guide children towards discovering or learning. The second sub-group, the Tried and Tested Methodologists have a more didactic model of teaching and learning: they see assessment as assessing what is taught. Both sub-groups resist criterion-referenced assessment, ie assessment in relation to Statements of Attainment. The Children's Needs Ideologists because it is in tension with the 'whole child' philosophy, the Tried and Tested Methodologists because it meant a radical change in behaviour for them. These teachers continued to incorporate effort or children's performance in relation to their background factors when making an assessment; their resistance to criterion-referenced approaches is epitomised by their reluctance to internalise or to have readily available the Statements of Attainment. As for the formative nature of teacher assessment, the Tried and Tested Methodologists were essentially summative: they would sit down at the end of a term or half-term and 'call up their memory' and record an assessment for each child in relation to each attainment target. The Children's Needs Ideologists would say that they were

constantly making formative assessments, but they could not articulate this, neither was it visible. In carrying out their essentially summative assessments, both of these sub-groups made use of assessment procedures with which they were familiar, such as the ILEA Check Points, their own or school-developed worksheets and tests and Maths worksheets from published schemes. This is in spite of the fact that the results from these did not relate to the levels and attainment targets of the National Curriculum.

The Evidence Gatherers collect evidence, written or drawn, in order to have a basis for making assessments. Some of these teachers collected hordes of evidence. At the end of each term or half-term, they would sit down and go through all the evidence and assign levels: this group does not rely on memory, since they feel that they need more than that to make an accurate assessment. However, often there is too much evidence to be used, and the teachers do not interrogate it all; part of the reason for collecting so much evidence seems to be that the evidence proves that the National Curriculum has been covered. In addition, collecting evidence in this way does not interfere with their normal teaching and classroom practice. These teachers tend to plan their work using the broad attainment targets and wait for assessment opportunities to arise rather than planning for assessment. The model of learning held by these teachers is essentially a traditional, didactic model: children learn what is taught and only what is taught; assessment follows teaching to check that the process is going according to plan. These teachers' view of criterion-referenced assessment is interesting in that they understand the idea of assessment in relation to criteria, but insist that context and pupil's background must sometimes be taken into account in judging performance; again they do not use Statements of Attainment. For this group of teachers, teacher assessment is essentially summative, however this group is becoming aware of a range of assessment procedures and recognised the importance of observation, and of children's talk, in making informal assessments.

Both the Evidence Gatherers and the Intuitives, rather than using SoAs, tended to have an overall notion of 'levelness' and therefore seemed to rely on implicit norms in judging children's performance. Some of the teachers, because of the quasi norm-referenced use of levels, tended to use Level 3 to indicate children of well above average attainment. Thus they ridiculed the possibility that children might, at this age, be reaching Level 4. Our observations, however, indicated that in some of the schools (and not always those in affluent areas) pupils were indeed able to achieve Level 4 in some parts of the curriculum.

Systematic Planners plan specifically for teacher assessment: they identify activities and tasks within their planned programme of teaching with specific Statements of Attainment in mind. They use multiple techniques for assessment: observation, open-ended questioning, teacher/pupil discussion, running records, scrutiny of written work. There are two sub-groups which we call Systematic Assessors and Systematic Integrators. The Systematic Assessors give daily, concentrated time to assessment and separate themselves off from the rest of the class to do it. For the systematic integrators, assessment is integrated with regular classroom work and often the teacher circulates through the class gathering her evidence in different ways. These teachers have a constructivist approach to learning: children learn in idiosyncratic ways and not always what is taught. They also have a particular view about assessment, which means that they are keen to arrive at shared meanings in relation to grading children's work with colleagues. This group do understand and operate a criterion-referenced model of assessment. They use Statements of Attainment openly and regularly, often broken down into more detailed 'Can-Do' lists. Information about effort, progress and performance in relation to background go into Records of Achievement or children's personal records. The significant difference between this group and the other two is their use of Statements of Attainment. These teachers seem to be carrying out formative assessment in that assessment feeds into their planning on a regular and systematic basis, the children's records are accessible and used (something which we did not see

with the other two groups) and they see real value in continuous, formative assessment as enhancing their professional development and effectiveness as teachers. This group of teachers do not necessarily maintain a model of formative assessment which involves making goals clear to the child, feeding back information directly related to those goals to the child, discussing and setting standards with the child and attempting to make them self-monitoring learners. In fact, this sort of feedback, in relation to specific National Curriculum goals, or assessment criteria, was almost never observed in the Key Stage One classes where we worked.

With no model of TA offered to teachers, it is perhaps not surprising that they came up with a range of approaches. These approaches were related to the teachers' views of teaching and learning, their general style of organisation and teaching, and their reaction to the imposition of National Curriculum Assessment. They were thus developing assessment practice in line with their general practice and philosophy of primary education. What is important though, for ensuring quality in teacher assessment, is that teachers should relate their assessment to the given criteria, or exemplars, and that they be encouraged to discuss the levels which they award to particular pieces of work and/or children. We believe that these models are helpful in allowing us to see where primary teachers may be in their views about assessment in relation to the use of criteria and exemplars, since we see these as key issues for ensuring quality in teacher assessment, particularly where it is to be used for reporting purposes.

Reliability and Validity in Teacher Assessment

A highly reliable test is of little use if it is not valid - but a test cannot be valid, in classical test theory, if it does not have a basic level of reliability. Although texts on educational measurement tend to maintain that validity is more important than reliability, in fact developments in psychological and standardised testing have

emphasised reliability. In the attempt to achieve highly accurate and replicable testing, the validity of the tests has often been neglected. The move towards performance-based assessment and the development of school-based teacher assessment are part of an attempt to redress the balance between reliability and validity. What is needed, is, of course, an appropriate balance between the two because they are in tension; Harlen's chapter argues this very cogently and puts forward our concept of quality in assessment, which derives from optimising validity and reliability.

In considering the traditional requirements for reliability and validity, Sadler suggests that, in view of the purpose of formative assessment, we reverse the polarity of the terms. In summative assessment reliability is presented:

> 'as a precondition for a consideration of validity. In discussing formative assessment, however, the relation between reliability and validity is more appropriately stated as follows: validity is a sufficient but not necessary condition for reliability. Attention to the validity of judgements about individual pieces of work should take precedence over attention to reliability of grading in any context where the emphasis is on diagnosis and improvement. Reliability will follow as a corollary.'
>
> (Sadler 1989 p122, my emphasis).

The requirement that students improve as a result of feedback can be seen as a consequential validity criterion for formative assessment. In this model the teacher must involve the student in discussion of the evaluation and what is needed to improve, otherwise the student is unlikely to be able to improve her work, furthermore the student needs to be involved in this process in order to shift to a process of self-monitoring. Formative assessment thus needs to demonstrate formative validity and in Sadlers's definition must involve feedback to the pupil; her involvement in and understanding of this feedback is crucial otherwise improvement is unlikely to occur.

We need to consider here the issue of purpose (and fitness for purpose). If assessment is to be used for certification or accountability then it needs an adequate level of reliability, in terms of consistency of performance and scoring, for comparability purposes. If however, the assessment is to be used for formative purposes, validity (content, construct and consequential aspects) is highly important and reliability is less so. Where teacher, school-based, assessment is concerned confusion often arises since reliability may be thought to be less important in a generic sense. However, this ignores the interaction with purpose : if teacher assessment is part of an accountability or certificating process, then reliability is important. The key, as Harlen makes clear, is how to achieve optimum reliability for the assessment's purpose while maintaining high validity.

Various methods of moderation together with training and the setting of criteria for grading are capable of enhancing reliability in teacher assessment. Mislevy suggests that moderation should be viewed as a way to specify the rules of the game, 'It can yield an agreed-upon way of comparing students who differ quantitatively, but it doesn't make information from tests that aren't built to measure the same thing function as if they did.' (Mislevy p72, 1992) The important point to emphasise is that the enhanced validity offered by teacher assessments is gained at a cost to consistency and comparability. Moderation is the process of attempting to enhance reliability which in technical terms can never be as great as in highly standardised procedures with all pupils taking the same specified tasks.

Enhancing Reliability

Where students perform the 'same' task for internal assessment purposes (e.g. a practical maths or science task or an essay with a given title) there are bound to be questions about the comparability of the judgements made by different teachers. Where there is no common task but common assessment criteria or common standards the

problem is different but the question the same: can we assume that the assessments are comparable across teachers and institutions?

Quality assurance is an approach that aims for standardisation or consistency of approach ie it focuses on the process of assessment. Quality control on the other hand focuses on ensuring that the outcomes are judged in a comparable way. Generally these two processes, and others which attempt to support comparability, are termed, in the UK, moderation. I shall now review the most relevant procedure from those outlined in chapter one, to explore how it can enhance teacher assessment.

Group moderation

This refers to the moderation of teachers' assessments by the common or consensus judgements of a group or panel of teachers and/or experts or moderators (SSABSA 1988). This is called variously group, consensus or social moderation, agreement panels or agreement trials.

In this chapter I will use the term group moderation. The key point is that it relies solely on teachers' professional judgement and is essentially concerned with quality assurance and the professional development of teachers, although it may serve only a quality control purpose.

In group moderation examples of work are discussed by groups of teachers or lecturers; the purpose is to arrive at shared understandings of the criteria in operation and thus both the processes and the products of assessment are considered. The process can be widened to groups of schools within a district or county: samples of graded work can be brought by one or two teachers from each school to be moderated at the district/county level. This will reveal any discrepancies among the various local groups and the same process of discussion and comparison would lead to some assessments

being changed in the same way as at the local level meeting. The teachers then take this information back to their own schools and discuss it in order to achieve a broader consensus.

Meetings across schools (as proposed for the English NC assessment programme in the TGAT Report, DES 1988) serve to enhance the consistency of judgements at the system level. They are, of course, more costly than meetings within a school/institution, but need to be evaluated in terms of their potential for supporting professional development for teachers particularly in relation to the processes of assessment, what counts achievement and how it may be best produced. Through discussion the assessments assigned to some pieces of work will be changed: 'The emphasis is on collegial support and the movement toward consensus judgements through social interaction and staff development' (Linn 1992 p25).

> 'In the use of social "(ie group)" moderation, the comparability of scores assigned depends substantially on the development of a consensus among professionals. The process of verification of a sample of student papers or other products at successively higher levels in the system (e.g. school, district, state, nation) provides a means of broadening the consensus across the boundaries of individual classrooms or schools. It also serves an audit function that is likely to be an essential element in gaining public acceptance'
>
> (p26 Linn op cit)

A pre-requisite to this process, of course, is a common marking scheme or a shared understanding of assessment criteria (i.e. the SoA in NCAss). The provision of exemplars, samples of marked or graded work, is sometimes a part of this process and, whilst not doing away with the need to have discussions about levels of performance, does aid teachers in getting an understanding of the overall standards. In National Curriculum Assessment at KSI in 1991 and 1992 when teachers were given little guidance in how to make their own assessments against the SoA, they found Children's

Work Assessed (SEAC 1993) booklets helpful in deciding what counted as evidence of performance for the different SoA (Source: NAPS data).

It is important to emphasise that it is not sufficient to focus on consistency of standards in marking or grading. Consistency of standards relates to ensuring that different teachers interpret the assessment criteria in the same way. However, when assessment tasks are open-ended, or not specified at all, then it is important to ensure consistency of approach: the assessment task or activity which is used and the way in which such tasks are presented to the pupil, or indeed contextualised, can affect performance quite markedly. To ensure consistency of approach, therefore, we need to ensure that teachers understand fully the constructs which they are assessing (and therefore what sort of tasks to set); how to get at the pupil's knowledge and understanding (and therefore what sort of questions to ask); and how to elicit the pupil's best performance (the physical, social and intellectual context in which the assessment takes place).

Group moderation is a key element of internal assessment, not only in terms of improving inter-marker reliability, but to support the process of assessment too. If we wish to be able to 'warrant assessment-based conclusions' without resorting to highly standardised procedures with all that this implies for poor validity, then we must ensure that teachers have common understandings of the criterion performance and the circumstances and contexts which elicit best performance.

The disadvantage of group moderation is that it is time consuming and costly and this may then be seen to add to any unmanageability in an assessment program. Its great advantage on the other hand lies in its effect on teachers' practice (Torrance 1982; Linn 1992; Radnor & Shaw 1994). It has been found that where teachers come together to discuss performance standards, or criteria, the moderation process becomes a process of teacher development with wash-back on teaching. It seems that coming together to discuss performance or scoring is less personally and professionally threatening than

discussing, for example, pedagogy. But discussion of assessment does not end there: issues of production of work follow on and this broadens the scope of discussion and impacts on teaching.

Moderation of Teacher Assessment in National Assessment at Key Stage One

In the section on moderation in the TGAT Report (DES 1988) the blueprint for the National Curriculum Assessment program, the authors argue for group moderation as the most appropriate method of moderation for NC Assessment because of its emphasis on communication and its ability to value and enhance teachers' professional judgements. However, the detailed account given of given of how such group moderation must work (paras 73 to 77) makes it clear that the process intended by TGAT is much closer to a scaling process, using the external SAT results to adjust the distributions of teachers' assessments.

The procedure proposed was as follows: groups of teachers would meet and consider two sets of results for each element of the NC: their own ratings and the results on the national tests, both expressed in terms of distributions over the levels of the NC e.g. % at Levels 1, 2 and 3. The task of the group would be to explore any lack of match between the two distributions. 'The general aim would be to adjust the overall teacher rating results to match the overall results of the national tests;...' (para. 74). The group would then go on to consider any discrepancies for particular schools using samples of work and knowledge of the circumstances of schools. "The moderation group's aim would here be to arrive at a final distribution for each school or pupil group. In general this would be the distribution on the national tests, with the implication that teachers' ratings would need adjustment, but departures from this could be approved if the group as a whole could be convinced that they were justified in particular cases.' (para 75). While the Report did accept

that the process could be carried out without the need for a group meeting at all (by simply adjusting the distribution to agree with those of the national testing) it argued for the opportunity for teachers to discuss mismatches between internal and external assessments in terms of their interpretation of the national curriculum itself and the national assessment instruments.

Thus what was being suggested here was a group process in which, rather than teachers bringing together pieces of work and agreeing on a common standard on the basis of their own professional judgements, involved teachers learning to adjust their ratings in the light of the external test results, which are considered to be the absolute standard (except in very occasional situations). Whilst these 'professional deliberations have a valuable staff development function...' (para. 76) it hardly looks like an assessment programme which values the professional judgement of teachers. It is essentially a quality control approach which aims to have a quality assurance role in the longer term.

The reaction from the Schools Examination and Assessment Council (SEAC) to the TGAT approach to moderation was negative for four reasons: it would place too many demands on teachers; it would take too long; for some attainment targets there would be no test data; moderation in a criterion-referenced system should be focused on individuals' scores, rather than scaling the outcomes of groups of pupils (Daugherty 1994). As Daugherty points out, it was less clear what model of moderation should replace the TGAT one.

In National Curriculum Assessment Year 2 teachers are required to make an assessment of each seven year old pupil's level of attainment on levels 1-4 of the scale 1-10 in relation to the attainment targets of the core subjects. Teachers may make these assessments in any way they wish, but observation, regular informal assessment and keeping examples of work, are all encouraged. In the first half of the

summer term and the second half of the spring term the pupils are given, by their teacher, a series of standard assessment tasks (SATs)* covering a sample of the core attainment targets.

As James & Conner (1993) point out, the SEAC Handbook for moderators emphasised consistency of approach (to conducting the assessments) and consistency of standards (inter rater reliability) which were to be achieved in 1991 and 1992 through the moderation process. In the event at Key Stage 1 a form of moderation by inspection was employed, for Teacher Assessment and SATs.

Of major concern in relation to reliability was that the statements of attainment (the assessment criteria specified in the curriculum documents) are not always sufficiently clear to allow teachers to make unambiguous judgements about performance; the criteria in this criterion-referenced assessment system were not specific enough for assessment purposes. In some of our research schools, which we describe as analytic, teachers discussed criteria and standards of performance among themselves and in these schools it is likely that assessments were more standardised and more comparable across classes than in other schools (Gipps, 1992b), a finding supported by the official evaluation in 1992 (NFER/BGC 1992). In the schools where discussion did take place it was partly because of the woolliness of the assessment criteria that these discussions were started. The visiting moderator helped in these discussions in some schools (James & Conner, 1993 op cit).

Owing to the problems with the statements of attainment there has been some concern over inter-rater (or judge) reliability. The technical evaluations carried out in 1991 indicate that SoAs were indeed interpreted differently by different teachers (NFER/BGC 1991) and that assessments made of the same attainment target by teacher assessment, SAT and an alternative test had unacceptable levels of variation

* In future SATs are to be called Standard Tasks at age 7, and Standard Tests at 11 & 14.

(Shorrocks et al 1992; Harlen 1993). The 1992 evaluation (NFER/BGC 1992) found that the match between TA levels and SAT levels was significantly greater in the second year of the assessment. A range of factors could be causing this, one at least being an artefact of the system rather than necessarily being due to teachers' changing assessment skills. In 1992 teachers did not have to commit themselves to their TA levels until after the SATs were done; it is possible then that the teachers' own assessments were affected by the SAT results; it is also possible that some teachers did not make a separate TA but simply used the SAT result where an attainment target was assessed by both.

The evidence on inter-rater reliability is limited in the UK other than the comparison of TA and SAT level, which is of dubious value. Furthermore the supervising body SEAC, the Schools Examination and Assessment Council, has admitted that there were good reasons for TA and SAT results not to align. TA, although less standardised, covers a wider range of attainments over a longer period of time, it may be less accurate than SAT assessment but is more thorough and offers a better description of overall attainment. 'The two forms of assessment should not therefore be regarded as identical' (p34 SEAC 1991); the determination of mastery was also an issue in 1991 and 1992: for the SATs all but one SoA had to be achieved to gain a particular level while in TA there was no such rule, and we do not know how teachers made their mastery decisions.

Evidence on inter-rater reliability of the SATs is therefore patchy but there is some evidence (James and Conner 1993; NFER/BGC 1992) which our case studies would support that teachers in schools who have, or make, the opportunity to discuss standards of performance, i.e. engage in group moderation, are developing common standards for assessment. Furthermore, the process of moderation had forced teachers to interact, negotiate meaning for SoAs, standardise judgements about

We need to move the debate beyond this assertion, since consistency of standards is not the prime requirement if the cost is to validity and teachers' professional involvement. We need assessment approaches which balance validity and consistency of standards; teacher assessment properly moderated, as described in this paper, can achieve this and enhance teachers' professional involvement and skills.

Conclusion

Assessment internal to the school in which the teacher is centrally involved is more professionally rewarding (in terms of enhancing teaching and learning) and valid (because of the range of skills and processes which may be included and the range of contexts in which assessment may take place) than external assessment, in which the teacher has little involvement. If, however, such assessment is to be used outside the classroom in reporting to parents or for accountability and certificating purposes, there must be some assurance to those receiving and using the results that there is comparability across teachers, tasks and pupils.

It is possible to ensure this through forms of statistical moderation, inspection of marked work by post and other quality control mechanisms. However, in line with the professional aspect of internal assessment, we advocate forms of moderation which are based on quality assurance and result in teacher development and enhanced understanding of the subject matter and its assessment. I have therefore concentrated in this paper on group moderation while James' paper describes discussion with a visiting moderator in which the teachers themselves are centrally involved. Group moderation, which involves discussing criteria as well as pieces of work, what counts as achievement and how such achievement is produced, is the most thorough of the quality assurance approaches. The considerable time (and cost) involved should not be underestimated, but can be seen as a valuable aspect of professional development.

individual children and discuss 'levelness' (Brown et al 1993). Concern about wider, national, levels of consistency remain however.

The process of group moderation, in which groups of teachers with or without a `moderator', or external expert, come together and discuss pieces of work or what counts as performance, greatly aids comparability. In some schools this process was going on but it needs to be supported and routinised if it is to have any serious impact on teachers' assessments.

From 1993 the process is to be called 'auditing', the term moderation having been dropped (DfE, 1993). The key difference is that rather than offering a system which supports moderation of the process and procedure of the assessments, evidence will be required that results conform to national standards: head teachers will have to ensure that teachers become familiar with national standards and keep evidence of assessment and records for audit when required. It is therefore a process of quality control rather than assurance.

The report by Sir Ron Dearing on the National Curriculum and Assessment recognises the role of teacher assessment both for formative purposes and, when moderated, for summative purposes, and recommends giving equal standing to TA and national tests in reporting to parents. The moderation process proposed in the interim report is, however, for a form of statistical moderation with national test results providing the consistency of standards against which to judge TA. National tests will:

> 'iii) provide a means of moderating teacher assessment in the subject so that discrepancies between the outcome of tests and teacher assessment can be investigated <u>in order to improve teacher assessments.</u>'
>
> (Dearing 1993, my emphasis)

In the same vein, he claims that moderation by groups of teachers or through visitation 'cannot readily produce the same consistency of standards as national tests' . (p50 op cit)

Finally, assuring quality through focussing on the processes of assessment and the assessment tasks, will I believe lead ipso facto to quality control of the outcomes of assessment; this together with an emphasis on validity will lead to confidence in comparability and high quality assessment.

References

Brown M, Gipps C, McCallum B (1993) The impact and use of national assessment results, Paper present to BERA conference Liverpool, September 1993.

Daugherty R (1994) National Curriculum Assessment: a review of policy 1988-1993 Falmer, in press

Dearing, Sir R (1993) The National Curriculum and its Assessment Interim report, July 1993 NCC and SEAC

Department for Education (1993) The Education (Assessment Arrangements for the Core Subjects) (Key Stage 1) Order 1993 Circular 11/93 DFE

DES (1988) National Curriculum Task Group on Assessment and Testing - A Report DES/WO

Gipps C (1992a) What we know about effective primary teaching, London File Tufnell Press

Gipps C (1992b) National Testing at Seven: What can it tell us?, Paper presented at AERA Conference 1992, San Francisco

Gipps C (1994) Beyond Testing: towards a theory of educational assessment Falmer, in press

Harlen W (1993) Quality Assurance and Quality Control in Assessment, BERA Assessment PTG

James M and Conner C (1993) 'Are Reliability and Validity Achievable in National Curriculum Assessment? Some Observations on Moderation at Key Stage One in 1992' The Curriculum Journal Vol 4 No 1

Linn R L (1992) Linking Results of Distinct Assessments Unpublished, CRESST, UCLA August

McCallum, E, McAlister S, Brown M & Gipps C (1993) 'Teacher Assessment at Key Stage One' Research Papers in Education Vol 8 No 3 pp305-27

Mislevy R J (1992) Linking Educational Assessments. Concepts, Issues, Methods and Prospects, Educational Testing Services, Princeton

NFER/BGC (1991) An Evaluation of National Curriculum Assessment. Report 3, June 1991

NFER/BGC (1992) An Evaluation of the 1992 National Curriculum Assessment at ICSI, September 1992

NISEAC (1991) Pupil Assessment in Northern Ireland Advice to Lord Belstead, Paymaster General, January

Radnor H and Shaw K (1994) 'Developing a Collaborative Approach to Moderation: the moderation and assessment project - south west' Chapter 6 in H Torrance (ed) Evaluating Authentic Assessment, Open University Press, Buckingham, in press

Sadler R (1989) 'Formative Assessment and the Design of Instructional Systems' Instructional Science 18 pp119-44

SEAC (1991) National Curriculum Assessment at Key Stage 3: a review of the 1991 pilots with implications for 1992, EMU: SEAC

SEAC (1993) School Assessment Folder

Senior Secondary Assessment Board of South Australia (SSABSA) (1988) Assessment and Moderation Policy Information Booklet No 2

Shorrocks D, Daniels S, Frobisher L, Nelson N, Waterson A and Bell J, (1992) ENCA 1 Project Report, London: SEAC

Torrance H (1982) Mode 3 Examining: six case studies Longman: Schools Council